DOCUMENT RESUME

ED 424 248 TM 029 120

AUTHOR Ross, John A.; Rolheiser, Carol; Hogaboam-Gray, Anne
TITLE Effects of Self-Evaluation Training on Narrative Writing.
SPONS AGENCY Social Sciences and Humanities Research Council of Canada, Ottawa (Ontario).; Ontario Dept. of Education, Toronto.
PUB DATE 1998-05-00
NOTE 24p.; Paper presented at the Annual Meeting of the Canadian Society for the Study of Education (Ottawa, Ontario, Canada, May 1998). Funding also provided by School District 13, Ontario.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criteria; *Elementary School Students; Foreign Countries; Intermediate Grades; *Narration; *Self Evaluation (Individuals); Tables (Data); *Training; *Writing (Composition); Writing Instruction
IDENTIFIERS *Authentic Assessment

ABSTRACT

Despite the burgeoning use of authentic assessment, few studies have examined effects on students. In this study, 148 students in 15 grade 4-6 classrooms were taught over an 8-week period how to evaluate their work. Their self-reflections were later compared with those of 148 control group students. Treatment group students became more accurate in their self-evaluations than controls. Contrary to the beliefs of many students, parents, and teachers, students' propensity to inflate grades decreased when teachers shared assessment responsibility. Treatment students also outperformed controls on narrative writing but the overall effect was small (ES=0.18). Poorer writers improved their writing much more if they were in the treatment rather than the control group (ES=0.58). The results of the treatment are attributed to the focusing effects of joint criteria development and use, and to the heightened meaningfulness of self-evaluation over other assessment data. An appendix presents the assessment scales for the Junior Division Narrative tests. (Contains 6 tables and 52 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

Effects of Self-Evaluation Training on Narrative Writing[1]

John A. Ross
Carol Rolheiser
Anne Hogaboam-Gray
Ontario Institute for Studies in Education

ED 424 248

Despite the burgeoning use of authentic assessment, few studies have examined effects on students. In this study, 148 students in 15 grade 4-6 classrooms were taught over an 8-week period how to evaluate their work (control N=148). Treatment group students became more accurate in their self-evaluations than controls. Contrary to the beliefs of many students, parents and teachers, students' propensity to inflate grades decreased when teachers shared assessment responsibility. Treatment students also outperformed controls on narrative writing but the overall effect was small (ES=.18). Poorer writers improved their writing much more if they were in the treatment than the control group (ES=.58). The results of the treatment were attributed to the focusing effect of joint criteria development and use, and to the heightened meaningfulness of self-evaluation over other assessment data.

Student appraisal practices in Ontario have shifted from an exclusive reliance on testing toward a more balanced approach in which classroom tests and examinations are supplemented with alternate forms such as portfolio assessment, performance evaluation, and self-evaluation. Assessment is more closely integrated with instruction, instruments and procedures are demystified, assessment is a continuous process rather than a terminal event, and teachers share authority with students.

Although self-evaluation has been implemented extensively in elementary schools, few systematic attempts to teach students how to evaluate their work have been reported and little is known about the effects of self-evaluation training on students' achievement. In this study we implemented an in-service program that provided a small sample of teachers with instruments and procedures for teaching grade 4-6 students how to evaluate their performance in narrative writing and measured the effects of the in-service on the accuracy of students' self-appraisals and the quality of their narrative writing.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

John Ross

TM029120

---

[1] Paper presented at the annual meeting of the Canadian Society for the Study of Education, Ottawa, May, 1998. The research was funded by the Social Sciences and Humanities Research Council of Council, the Ontario Ministry of Education, and School District 13 (formerly the Durham Board of Education). The views expressed in the report do not necessarily reflect the views of the Council, the Ministry of the school district. Corresponding author: Dr. John A. Ross, Professor & Head, OISE/UT Trent Valley Centre, Box 719, 150 O'Carroll Ave., Peterborough, ON K9J 7A1

Theoretical Framework

Motivation for Authentic Assessment Practices

Proponents of authentic assessment practices, defined as assessment that "occurs in motivating contexts with meaningful tasks that are a part of daily instruction" (van Kraayenoord & Paris, 1993, p. 524), claim that a shift from traditional to authentic assessment will increase student achievement (e.g., Stiggins, 1994; Wiggins, 1993). Learning is enhanced because authentic assessment encourages teachers to focus on the objectives to be measured and the assessments provide teachers with more accurate information than traditional tests, enabling teachers to respond more precisely to students' learning needs.

There is consistent, though not extensive, evidence that authentic assessment influences teachers' instructional practices in positive ways (e.g., Dorfman, 1997; Khattri, Kane, & Reeve, 1995; Koretz, Stecher, Klein, & McCaffrey, 1994; Lipson & Mosenthal, 1997). Few studies have asked whether these changes increase student achievement. Studies of authentic assessment that attempted to answer questions about student outcomes produced mixed results. For example, Shepard, Flexer, Hierbert, Marion, Mayfield, and Weston (1996) found that performance assessment had a small positive effect on student achievement in mathematics but not reading. Bangert (1997) found that peer assessment made a large contribution to student learning in a graduate statistics course. A state-wide portfolio assessment program was abandoned, partly because student scores on traditional and alternate evaluations declined (Chrispeels, 1997). There is some evidence that students' study habits change when performance assessment (Lee & Suen, 1995) or portfolio assessment (Slater, Ryan, & Samson, 1997) is introduced.

Rationale for Linking Self-Evaluation to Achievement

Our expectation that a self-evaluation assessment system would enhance student achievement was based on four arguments. Students will learn more because (i) self-evaluation will focus student attention on the objectives measured, (ii) the assessment provides teachers with information they would otherwise lack, (iii) students will pay more attention to the assessment, and (iv) student motivation will be enhanced.

Self-evaluation focuses student attention. It has long been demonstrated that being clear about goals makes a positive contribution to performance (Locke, Shaw, Saari, & Latham, 1981). Several studies (reviewed in Hillocks, 1986) found that the quality of student writing improved when teachers explained the criteria on which student work would be judged. If students apply these criteria to assess their work, the effects should be even stronger (through rehearsal and focusing). Students should develop a clearer understanding of what they are supposed to do and how well they are doing it. In addition, students report that they have a better grasp of academic expectations when they are involved in setting the criteria on which their work will be judged (Ross, Rolheiser, & Hogaboam-Gray, in press-c).

These benefits are likely to accrue only if students are provided with outcome-based criteria at an appropriate level of generality. Not all self-evaluation procedures recommended for classroom use do so. For example, many of the instruments in Rhodes (1993) direct student attention to processes not outcomes: Students ask themselves whether they used each of the steps in the Writing Process, rather than how well they wrote. In contrast, Kulm (1994) recommends that teachers begin self-evaluation by involving students in the construction of scoring rubrics which students then use to appraise their work. But even the joint construction of rubrics might be insufficient to focus student attention if the rubrics are too task-specific (making it difficult for students to detect the underlying learning objectives), too general (simply an array of superlatives that fail to indicate what is essential in a quality response), or too complicated for students to use easily.

Self-evaluation provides teachers with information otherwise unobtainable. Conventional test procedures and many authentic performance tasks provide no information about students' inner states during task performance, their subsequent interpretations about the quality of their work, and the goals they set in response to feedback. Self-evaluation is unique in asking students to reflect on their performance. Self-evaluation instruments that elicit information about students' effort, persistence, goal orientations, attributions for success and failure, and beliefs about their competence give teachers a fuller understanding of why students performed as they did. When incorporated into teachers' deliberative planning, data generated from self-evaluations enables teachers to present content and anticipate impediments to learning, especially motivational obstacles. But many self-evaluation procedures, particularly those consisting of closed-response instruments, fail to provide such rich detail.

Students pay more attention to self-evaluation than to other assessments. As students move through the school system their skepticism about the validity of test scores increases (Paris, Lawton, Turner, & Roth, 1991), a trend that has also been observed in portfolio assessment projects (Paris, Turner, Muchmore, Perry, 1995). Interviews with grade 5-11 students indicated that students viewed self-evaluation more positively than other kinds of assessment (Ross et al., in press-c). Students liked self-evaluation because it increased clarity about expectations, was fairer (because it enabled students to include in their summative evaluation the effort they put into the task), and gave students feedback they could use to improve the quality of their work. Some students reported that self-evaluation was more useful to them than feedback from the teacher because with teacher feedback they focused on what they did well or on the grade, whereas with self-evaluation they focused on what they needed to work on. Students also had a variety of negative feelings and beliefs about evaluation, for example, that it enabled the undeserving to give themselves a higher grade, that some students lacked the expertise to mark their work, and that in some classes self-evaluation was counted only if it was confirmed by the teacher's appraisal. We also found that student attitudes to self-evaluation became more positive with experience. If students are paying more attention to evaluation data, student achievement should increase. But the benefits are not likely to accrue if the rubrics on which the assessment is based are inappropriate or covert (as argued above). In addition students need help in using evaluation data to set goals. The provision of valid performance data without goal setting support

can decrease achievement for lower ability students (Fuchs, Fuchs, Karns, Hamlett, Katzaroff, & Dutka, 1997).

Student motivation will be enhanced. In addition to self-evaluation influencing achievement directly (as described above), self-evaluation has an indirect effect through self-efficacy. The most powerful contributor to higher self-efficacy is mastery experience (Bandura, 1986), that is, students who have been successful in the past anticipate they will be successful in the future. But even unsatisfactory performance might not lead to depressed confidence if the student believes that he or she could be successful by adopting a different strategy (Schunk, 1995). What is crucial is how the student evaluates the performance. Higher self-efficacy translates into higher achievement (Pajares, 1996). Positive self-evaluations encourage students to set higher goals and commit more personal resources to learning tasks (Bandura, 1986; Schunk, 1995). Negative self-evaluations lead students to embrace goal orientations that conflict with learning, select personal goals that are unrealistic, adopt learning strategies which are ineffective, exert low effort and make excuses for performance (Stipek, Recchia & McClintic, 1992). Wagner (1991), in a linear structural model, found positive path coefficients from self-evaluation to self-efficacy and from self-efficacy to performance.

Student Skill in Evaluating Their Work

These arguments suggest that self-evaluation is potentially a powerful stimulant of achievement. But in our interviews with students we found that they harbored a variety of misconceptions about the process, for example, many did not appreciate the role that evidence plays in self-evaluation (Ross et al., in press-c) or how discrepancies between student and teacher appraisals are resolved (Ross, Rolheiser, & Hogaboam-Gray, in press-b). Self-evaluation is unlikely to have a positive impact on achievement if these misconceptions are not addressed by teaching students how to assess their work.

Students' self-evaluations tend be inflated. Accuracy is a matter of degree. The self-evaluations of even young children correlate reasonably well with their teachers' appraisals when students are asked to make a global assessments, comparing their ability to that of their classmates (Crocker & Cheeseman, 1988). Accuracy is much lower for specific tasks, even for adults, if information about a specific ability is lacking or difficult to process (Bandura, 1977). Elementary students tend to over-estimate their success on school tasks, in part because they expect that teachers will give them tasks that they can complete (Schunk, 1996), but also because young students lack the cognitive skills required to integrate information about their abilities and they are more vulnerable to wishful thinking. Overestimates of specific performance are likely to lead to complacency and reduced effort. For example, the child who does not recognize the need for help will not seek it (Markman, 1979). Simply requiring self-evaluation is unlikely to have an effect on achievement. Students have to be taught how to evaluate their work accurately.

Results of Previous Attempts to Teach Self-evaluation

Few studies have examined the effects of teaching students how to self-evaluate in classroom settings over a sustained period (i.e., four weeks or more). The results have been mixed, with positive outcomes reported for achievement by Fontana and Fernandes (1994) and Sparks (1991) but not Aichison (1995). Similarly Ross (1995a) but not Connell, Carta, and Baer (1993) reported positive effects of self-evaluation on students' learning strategies.

There have also been studies in which evaluation of writing skills has been embedded in a broader instructional program. Hillocks (1986) reviewed seven studies (all but one were unpublished dissertations) in which students were given scales for judging writing samples. Students used the scales to assess the writing of their peers, to give editing suggestions, and to re-write deficient passages. In some instances they evaluated their own writing, although in Hillocks' review self-evaluation was a minor theme in most of these studies.

The only study of self-evaluation in the language area (Arter, Spandel, Culham, & Pollard, 1994) gave grade 5 students direct instruction on the meaning of six traits of essay writing. The teacher, without student participation, determined the traits. Students scored a sample of essays and applied trait analysis to their own writing over a five-month period. The treatment group outperformed controls on one of the six traits (ideas). But the analysis procedures failed to protect the findings from Type I error. The authors should have used multivariate analysis (to deal with the possibility of multicollinearity among the six dependent variables) or applied a Bonferroni adjustment (Serlin, 1993). When the latter was applied (the alpha of $p<.05$ was divided by six, the number of comparisons in the same set of dependent variables) none of the results was significant.

No studies of the effects on student accuracy of teaching self-evaluation in elementary and secondary classrooms have been reported. Research on university students indicates that accuracy improves when professors and students agree on assessment criteria (Falchikov & Boud, 1989) and when students are asked to justify their assessments (Boud, Churches, & Smith, 1986). There is also evidence from short duration lab studies that the self-evaluation accuracy of elementary students can be improved by influencing goal conditions (Butler, 1990) and drawing attention to previous performance (Stipek, Roberts, & Sanborn, 1984).

Research Questions and Predictions

Our approach to teaching students how to evaluate their work began in a study of the student assessment practices of exemplary cooperative learning teachers (Ross, Rolheiser, & Hogaboam-Gray, in press-a). We organized their strategies as a four-stage process: (i) involve students in defining evaluation criteria, (ii) teach students how to apply the criteria, (iii) give students feedback on their self-evaluations, and (iv) help students use evaluation data to develop action plans. Strategies for each stage were elaborated by a team of teachers and reported as a series of action research case studies and classroom usable tools (Rolheiser, 1996). Use of these strategies had a positive effect on student attitudes to evaluation in some but not all of the pilot test classrooms (Ross et al., in press-b; in press-c). Our goal in this study was to determine

whether teaching students how to evaluate their work would improve achievement in language (narrative writing) of students in grades 4-6. Our research questions and hypotheses were:

1. Will self-evaluation training increase the accuracy of students' self-assessments? We anticipated that students in the treatment group would evaluate their work more accurately because all four stages in our model reduce uncertainty about the criteria for judging academic work.

2. Will self-evaluation training contribute to language achievement? We anticipated that focusing student and teacher attention on performance criteria (Stages 1 and 2) would enhance achievement.

Method

Sample

Students in the classrooms of 15 volunteer grade 4-6 teachers, in a large school district, constituted the treatment group. They were matched with a student control group from a convenience sample of 15 volunteer teachers in an adjacent board. Within each class we randomly selected data from 10 students for analysis. In one treatment class only eight students obtained parental consent so we randomly deleted two students from one of the control classes. The total sample was 296 students.

Instruments

Students completed a battery of instruments at the beginning and end of the project in the following sequence: On day 1 they completed a survey (described below) consisting of a self-efficacy measure (how sure they were they could write a good short story) and a locus of control measure. On Day 2 they wrote a short story. On Day 3 they evaluated their short story, shared their attitudes toward self-evaluation and completed a goals orientation survey about their feelings when writing their short story.

Student Achievement. Students completed (a) a pre- and post-test narrative writing task. Teachers were asked to present the writing task in their usual manner. Teachers could have a class discussion of possible topics but they had to emphasize to students that it was an individual writing task. Teachers described the criteria on which the stories would be marked (plot or story development, characters, setting, providing interest for the reader and grammar/spelling). Students wrote a rough copy and then a final copy. We asked teachers not to provide editing help. There was little difference between rough and final copies, except in one treatment class in which the teacher edited some stories. In this class the rough copies were marked; in all other classes the final copies were coded.

We developed a six-level coding scheme (displayed in the appendix) by elaborating descriptions in the provincial writing rubric. Two anchor papers were identified for each level. Over a two-week period two teachers used the rubric to mark 592 stories. An English consultant who had

been a trainer in provincial writing assessment programs trained the markers (by reviewing the six levels in terms of the anchor papers, marking and discussing additional papers from the pilot test) in the first two days. The two markers then independently graded each paper, assigning a holistic 1-6 score. After each set of ten papers the markers resolved discrepancies in their assessments through discussion. Before discussion the between-rater agreement was Cohen's κ=.63 for perfect agreement and κ=.83 for agreement within one level on the six-point scale. The papers were marked in random order, intermingling pre- with post-tests and treatment with control students. The markers and the trainer were blind to the experimental conditions of the students and to study goals.

(b) Accuracy of self-evaluation was calculated from the achievement data and from student responses to survey items administered immediately after the achievement task. Students used a 1-10 scale (anchored by 1=not well and 10= very well) to rate the quality of their story. They used the same 1-10 scale to answer five additional probes to rate how well they wrote each part: plot, characters, setting, interest for the reader, and grammar and spelling. These six items were averaged to create a 1-10 mean score for each student. The self-evaluation scores and the achievement scores were bifurcated at their medians and combined to create three groups: accurate (low self-evaluation with low achievement or high self-evaluation with high achievement), underestimate (low self-evaluation with high achievement) and overestimate (high self-evaluation with low achievement).

Student Instruments for Estimating Sample Equivalence. Measures predicting student achievement in previous research were administered to estimate the pretest equivalence of the two groups. The goals orientation survey consisted of 15 items from Meece, Blumenfeld, and Hoyle (1988) distinguishing three orientations toward learning tasks: mastery (e.g., "The work made me want to find out more about the topic."), ego (e.g., "I wanted others to think I was smart."), and affiliative (e.g., "I wanted to help others with their work."). Students with a mastery orientation are more likely to be successful learners (Meece et al., 1988). Student self-efficacy consisted of 6 items identical to the self-evaluation measure except that each asked "how sure are you that you could..." rather than "how sure are you that you [did]". In previous research (Pajares, 1996) self-efficacy predicted student achievement. Attitudes to self-evaluation scale consisted of 10 Likert items adapted from Paris, Turner, and Lawton (1990) and Wiggins (1993). Although no previous studies have examined the relationship between achievement and attitudes to evaluation, we argued above that two are theoretically linked.

Teacher Instruments for Estimating Sample Equivalence

We also examined pretest equivalence by comparing the 15 teachers in each sample on constructs linked to assessment practice or language achievement. Ten Likert items measured teachers' use of assessment methods that are fair, transparent, participatory, and collaborative (e.g., "My students help me interpret assessment results."). Teachers also completed 16 items from Gibson and Dembo (1984) measuring personal teaching efficacy (e.g., "When I really try, I can get through to even the most difficult students.") and general teaching efficacy (e.g., "The amount that a student can learn is primarily related to family background."). Both types of teacher efficacy correlate with teachers' willingness to try new ideas and student achievement (evidence reviewed in Ross, 1995b). Teachers also provided demographic information (e.g., gender, experience,

certification). Although evidence of the effects of demographic variables on instructional practice are weak, higher student achievement has been linked to having a graduate degree (Ferguson, 1991).

Experimental Conditions

Teachers in the treatment condition attended 3 three-hour, after-school in-service sessions distributed over the eight weeks of the field test. During this time they attended four brief team meetings in their schools to review progress and solve problems that arose during their enactment of the treatment. The in-class activities consisted of 4-30 minute lessons in which the teacher demonstrated a particular self-evaluation technique (e.g., constructing a rubric for writing) or engaged students in a discussion of their self-evaluations. There were 12 short practice sessions in which students completed a 3-5 minute self-evaluation using a form provided by the teacher. The in-service sessions and the handbook (Rolheiser, 1996) provided examples of lessons and practice activities that teachers could adapt. Although some of these examples focused specifically on language development, most were focused on assessing social skills. Teachers had complete control over how they adapted these materials to the language curriculum.

During the 8 weeks of the project the control group teachers continued teaching language as they usually did, including self-evaluation if that was part of their practice, but not emphasizing it. Control group teachers (unlike the treatment group) received a half-day of additional prep time to work on their writing curriculum.

Analysis

Descriptive statistics (means, standard deviations, reliabilities) for all student and teacher variables were compiled. Prior to inferential statistics all variables were normalized using log transformations. Pretest equivalence of the treatment and control groups was determined through a series of t-tests in which the dependent variables were student and teacher variables associated with achievement or evaluation practice in previous research; the independent variable was experimental condition. For the first research question, the proportion of students with accurate self-appraisals at the beginning and end of the project in the treatment and control groups were compared in contingency tables, using chi-square to determine statistical significance. For the second research question, an analysis of covariance was conducted in which the dependent variable was post-test achievement, the covariate was pre-test achievement, and the independent variable was experimental condition.

Results

Teacher Data

Table 1 describes the teacher variables. The reliabilities of the three scales were acceptable. At the beginning of the project there were no significant differences between treatment and control group teachers in terms of their self-reported use of authentic assessment practices, personal

teaching efficacy, and general teaching efficacy. Table 1 also shows the teachers were similar in age, experience, gender and qualifications.

Table 1 About Here

Student Data

Table 2 summarizes the reliabilities (Cronbach's Alpha) of the student variables. The reliabilities for self-efficacy, self-evaluation, self-evaluation attitudes, and mastery goal orientations were adequate. The reliabilities for mastery and ego goal orientations were borderline.

Table 2 About Here

Table 3 summarizes the means and standard deviations of the student variables for each treatment condition. There was one pretest difference between the groups. Treatment students significantly outperformed control students on the pretest writing task [$t(290.164)=4.79$, $p=.022$], a concern because the pretest writing task strongly predicted posttest writing scores [$r=.605$, $p<.001$] and posttest accuracy [$r=.384$, $p<.001$]. There was no other pretest student difference between the groups.

Table 3 About Here

The first research question asked whether training in self-evaluation increased the accuracy of student appraisals. There were no significant differences between treatment and control group students in self-evaluation accuracy on the pretest [$\chi^2(1, 284)=2.992$, $p\leq.084$]. On the post-test, treatment group students were significantly more accurate in their self-assessments [$\chi^2(1, 277)=7.037$, $p<.008$]. Table 4 shows the posttest accuracy rate within each experimental condition for three groups: students who underestimated, overestimated, and accurately appraised their pretest performance. Although the trends are clear, none of the differences reached statistical significance. Very few students (less than 2% of the sample) underestimated their performance on the pretest. Of those who did, two of the four treatment students accurately evaluated their posttest story while the single control group student continued to underestimate [Fisher's exact test $p<.600$]. Students who were accurate in assessing their pretest story (30% of the total sample) were more likely to continue to be accurate if they were in the treatment than the control group [$\chi^2(1, 90)=2.960$, $p<.085$]. Students who overestimated their performance on the pretest (the largest group at 68% of the sample) were more likely to accurately evaluate their writing on the posttest if they were in the treatment than in the control group [$\chi^2(1, 202)=3.803$, $p\leq.051$]. Most of these students continued to overestimate their performance. The data in Table 4 suggest that the treatment had a positive impact on the accuracy of students' self-evaluations, although the effect was small and a substantial number of students continued to over-estimate their performance even after eight weeks of self-evaluation training.

Table 4 About Here

Table 5 shows the results of the analysis of covariance of student achievement. The dependent variable was posttest writing score, the independent variable was experimental condition, and the covariates were pretest writing score and self-evaluation accuracy. Table 5 shows that only one of the covariates, pretest writing score was a significant predictor of achievement. Students who scored high on the pretest also scored high on the posttest as expected by the correlation (r=.605). The table shows that the correlation of pretest accuracy with post achievement (-.245) was spurious. Pretest achievement, which predicted post achievement, was one of the terms in the calculation of pretest accuracy. When pretest scores were controlled, self-evaluation accuracy did not predict achievement.

Table 5 About Here

There was a main effect for experimental condition. Both groups improved over the 8 weeks of the study. There was a pre- to posttest gain of ES=.40 for the treatment group and ES=.27 for the control (in each case the posttest mean was subtracted from the pretest mean and divided by the standard deviation of the pretest). Table 5 shows that students who were taught how to evaluate their work wrote better narratives than students in the control group (after controlling for pretest differences between the groups). But the effect of the treatment was very small (treatment versus control ES=.18), accounting for only 2% of the variance compared to 28% for the pretest covariate.

Table 5 also shows there was a treatment X pretest interaction. To understand it, we divided the sample into low (pretest scores 1-3) and high (pretest scores 4-6) achievers. Inspection of the cell means (after posttest achievement scores had been transformed) shows, in Table 6, that the treatment had an impact only on low achievers. Students who produced poor writing samples on the pretest and were then taught how to evaluate their work substantially outperformed similar students in the control group (ES=.58). In contrast, students who wrote well on the pretest performed equally well on the posttest, regardless of whether they were given self-evaluation training.

Table 6 About Here

Discussion

The first finding was that teaching self-evaluation skills increased the accuracy of student self-appraisals. The greatest impact was for students who were overestimating their performance, a sizeable proportion of the sample. This is an important finding for two reasons. First, students are unlikely to change how they go about writing narratives nor are they likely to seek help from teachers and peers, if they believe their work meets classroom standards. Second, students are concerned about accuracy. Some students believe they lack the expertise to assess their work accurately; others anticipate that given an opportunity to have input to their grades, their peers will cheat (Ross et al., in press-c). Teachers have reservations about self-evaluation for the same reasons and teachers perceive parents to have similar fears (Ross et al.., in press-a). But in this study, when teachers shared assessment control with students the tendency to inflate grades

decreased. Student involvement in rubric construction and receiving feedback on their application of the criteria gave students a clearer understanding of classroom standards. In addition, students were required to talk about the reasons for their self-evaluations, especially when there were discrepancies between a self-evaluation and the teacher's judgment. This focus on rubric-grounded evidence reduced the influence of other bases that students could use to award themselves grades, such as amount of effort and self-aggrandizement, that students rely on when standards are unarticulated.

The size of the impact of training on accuracy was small. One possible explanation might be that students participated in the development of classroom writing rubrics but did not have input to the rubrics on which their posttest stories were graded. If they had seen the rubric used to mark their pre- and post-test stories, their judgments about how well they did might have been more accurate. Since the rubric was not available to them we do not know the basis for their self-evaluations. It could have been rubrics for writing they had co-produced in class, an intuitive comparison between the piece they wrote for the research and pieces they had written previously, or they may have been comparing their performance to the writing typically produced by their peers. An alternate explanation for the weak effects of the treatment on accuracy is that we focused on a learning objective that receives extensive instructional attention in every grade. Students might have been so knowledgeable about what counts in writing that the focus on criteria and evidence contributed little to their understanding of what they were supposed to do. It might be that self-evaluation training would have a greater impact if focused on less familiar learning objectives.

The second finding was that self-evaluation training had a positive effect on achievement but only among weaker writers. The overall effect of the treatment was small (ES=.18), below the average effect size (.28) for the 75 writing composition treatments reviewed in Hillocks' (1986) meta-analysis. Part of the explanation for small overall effects might be the duration of the treatment (8 weeks). Hillocks found that treatments of less than 17 weeks had a lower effect size (.21). It may be that a longer duration is required before students' misconceptions about self-evaluation are overcome. In addition students need time to accept the idea that they have a role in assessment. In students' prior school experience evaluation was the teacher's exclusive prerogative. Another explanation is that the materials teachers used were not exclusively focused on writing, although only writing performance was measured in the study. Teachers reported using self-evaluation for social skills(following examples in Rolheiser, 1996) and in other subjects. The effects of self-evaluation training on writing skills may have been diluted.

Self-evaluation had a much larger impact on the performance of students who wrote poorly at the beginning of the study (ES=.58). The reason might be that self-evaluation training gave poorer writers explicit feedback on what they needed to improve on that was more meaningful to them than the feedback they usually received from the teacher. In our previous studies of student cognitions about self-evaluation (Ross et al., in press-b; in press-c) students reported paying more attention to self-evaluation because they understood the criteria, they felt ownership of the data, and they felt empowered because the teacher trusted them to rate themselves fairly. But why did the treatment have minimal effect on higher achievers? It might

be that the rubrics were oriented toward lower performers. It might also be that better writers knew what was expected of them and did not need the criteria to be spelled out in rubrics (at least for narrative writing).

The overall effect of self-evaluation on student achievement in this study was greater than has been reported for other authentic assessment measures. For example, Shepard et al. (1996) found that a project to introduce performance assessments (grade 3 teachers had weekly workshops for a year) had no overall effect on reading achievement and only a small effect on mathematics (ES=.13). The data from this study, particularly regarding the performance of poorer writers, presents a more encouraging picture.

This study produced knowledge of two types. For researchers the study contributed evidence of the consequential validity of authentic assessment, a topic that has been seriously neglected despite recognition among test developers that the consequences of test use is a key factor in determining the worth of assessment instruments (Linn, 1997; Messick, 1995; Moss, 1992; Shepard, 1997). For teachers the study suggests that self-evaluation might be a useful mechanism for increasing student achievement and the accuracy of self-appraisal. Thoughtfully designed self-evaluation procedures that provide students with explicit criteria at an appropriate level of generality, that provide for student involvement in assessment decision making, that elicit student cognitions about their performance, which ground student goal setting in accurate data, and that are integrated with sensitive instruction may provide teachers with a powerful lever for enhancing student learning.

References

Aitchison, R. E. (1995). The effects of self-evaluation techniques on the musical performance, self-evaluation accuracy, motivation and self-esteem of middle school instrumental music students. Unpublished doctoral dissertation, University of Iowa.

Arter, J., Spandel, V., Culham, R., & Pollard, J. (1994, April). The impact of training students to be self-assessors of writing. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84, 191-215.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice-Hall.

Bangert, A. W. (1997, April). The instructional and motivational benefits of peer assessment. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Boud, D. J., Churches, A. E., & Smith, E. M. (1986). Student self-assessment in an engineering design course: An evaluation. International Journal of Applied Engineering Education, 2(2), 83-90.

Butler, R. (1990). The effects of mastery and competitive conditions on self-assessment at different ages. Child Development, 61, 201-210.

Chrispeels, J. H. (1997). Educational policy implementation in a shifting political climate:The California experience. American Educational Research Journal, 34(3), 453-481.

Connell, M. C., Carta, J. J., & Baer, D. M. (1993). Programming generalization of in-class transition skills: Teaching preschoolers with developmental delays to self-assess and recruit contingent teacher praise. Journal of Applied Behavioral Analysis, 26, 345-352.

Crocker, A. C., & Cheeseman, R. G. (1988). The ability of young children to rank themselves for academic ability. Educational Studies, 14(1), 105-110.

Dorfman, A. (1997, April). Teachers' understanding of performance assessment. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. Review of Educational Research, 59(4), 395-430.

Ferguson, R. (1991). Paying for public education: New evidence on how and why money matters. Harvard Educational Review, 465-498.

Fontana, D., & Fernandes, M. (1994). Improvements in math performance as a consequence of self-assessment in Portuguese primary school pupils. British Journal of Educational Psychology, 64, 407-417.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C., Katzaroff, M., & Dutka, S. (1997). Effects of task-focused goals on low-achieving students with and without learning disabilities. American Educational Research Journal, 34(3), 513-543.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. Journal of Educational Psychology, 76(4), 569-582.

Hillocks, G. (1986). Research on written composition: New directions for teaching. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills.

Khattri, N., Kane, M., & Reeve, A. (1995). How performance assessments affect teaching and learning. Educational Leadership, 53(3), 80-83.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. Educational Measurement: Issues and Practice, 13(3), 5-16.

Kulm, G. (1994). Mathematics assessment: What works in the classroom. San Francisco: Jossey-Bass.

Lee, S., & Suen, H. (1995, April). Effect of two assessment techniques on study strategy use. Paper presented at the annual conference of the American Educational Research Association, San Francisco.

Linn, R. (1997). Evaluating the validity of assessments: The consequences of use. Educational Measurement: Issues and Practice, 16(2), 14-16.

Lipson, M. Y., & Mosenthal, J. (1997, March). The differential impact of Vermont's writing portfolio assessment on classroom instruction. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Locke, E., Shaw, K., Saari, L., & Latham, G. (1981). Goal setting and task prformance: 1969-1980. Psychological Bulletin, 90(1), 125-152.

Markman, E. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. Child Development, 50, 643-655.

Meece, J., Blumenfeld, P., & Hoyle, R. (1988). Students' goal orientations and cognitive engagement in classroom activities. Journal of Educational Psychology, 80(4), 514-523.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5-8.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62(3), 229-258.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. Review of Educational Research, 66(4), 543-578.

Paris, S., Lawton, T., Turner, J., & Roth, J. (1991). A developmental perspective on standardized achievement testing. Educational Researcher, 20(5), 12-20, 40.

Paris, S., Turner, J., & Lawton, T. (1990). Students' views of standardized achievement tests. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Paris, S. G., Turner, J. C., Muchmore, J., & Perry, N. (1995). Teachers' and students' perceptions of portfolios. Journal of Cognitive Education, 5(1), 6-40.

Rhodes, L. K. (Ed.). (1993). Literacy assessment: A handbook of instruments. Portsmouth, NH: Heinemann.

Rolheiser, C. (Ed.). (1996). Self-evaluation...Helping kids get better at it: A teacher's resource book. Toronto: OISE/UT.

Ross, J. A. (1995a). Effects of feedback on student behavior in cooperative learning groups in a grade 7 math class. Elementary School Journal, 96(2), 125-143.

Ross, J. A. (1995b). Strategies for enhancing teachers' beliefs in their effectiveness: Research on a school improvement hypothesis. Teachers College Record, 97(2), 227-251.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (in press-a). Student evaluation in cooperative learning: Teacher cognitions. Teachers and Teaching.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (in press-b). Teaching students how to self-evaluate when learning cooperatively: The effects of collaborative action research on teacher practice. Elementary School Journal.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (in press-c). Skills training versus action research in-service: Impact on student attitudes to self-evaluation. Teaching and Teacher Education.

Schunk, D. H. (1996, April). Self-efficacy for learning and performance. Paper presented at the annual meeting of the American Education Research Association, New York.

Schunk, D. H. (1995, April). Goal and self-evaluative influences during children's mathematical skill acquisition. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Serlin, R. (1993). Confidence intervals and the scientific method: A case for Holm on the range. Journal of Experimental Education, 61(4), 350-360.

Shepard, L. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice, 16(2), 5-8,13,24.

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1996). Effects of introducing classroom performance assessments on student learning. Educational Measurement: Issues and Practice, 15(3), 7-18.

Slater, T., Ryan, J., & Samson, S. (1997). Impact and dynamics of portfolio assessment and traditional assessment in a college physics course. Journal of Research in Science Teaching, 34(3), 255-271.

Sparks, G. E. (1991). The effect of self-evaluation on musical achievement, attentiveness and attitudes of elementary school instrument students. Unpublished doctoral dissertation, Louisiana State University and Agricultural and Mechanical College.

Stiggins, R. (1994). Student-centered classroom assessment. New York: Merrill.

Stipek, D., Recchia, S., & McClintic, S. (1992). Self-evaluation in young children. Monographs of the Society for Research in Child Development, 57, 1-84.

Stipek, D., Roberts, T., & Sanborn, M. (1984). Preschool-age children's performance expectations for themselves and another child as a function of the incentive value of success and the salience of past performance. Child Development, 59, 1983-1989.

van Kraayenoord, C., & Paris, S. (1993, December). Self-assessment: Children's perspectives on their literacy activities in the classroom. Paper presented at the National Reading Conference, Charleston, SC.

Wagner, M. J. (1991). Goal-setting, self-efficacy, and learning performance: A causal model. Unpublished doctoral dissertation, University of Arizona, AAC 911978 ProQuest - Dissertation Abstract.

Wiggins, G. (1993). Assessing student performance: Explore the purpose and limits of testing. San Francisco, CA: Jossey-Bass.

Table 1: Means, Standard Deviations, and Reliabilities of Teacher Variables

				Treatment (N=15)		Control (N=15)	
	Number of Items	Range	Alpha	Mean	SD	Mean	SD
Assessment Practices	10	1-6	.84	3.88	.66	3.79	.97
Personal Teaching Efficacy	9	1-6	.69	4.18	.49	4.30	.50
General Teaching Efficacy	6	1-6	.77	3.99	.64	4.03	1.02
Mean years of teaching				13.40		16.07	
Gender							
male				5		5	
female				10		10	
Grade of class							
3/4				1		0	
4				3		4	
4/5				0		2	
5				5		3	
5/6				3		2	
6				3		4	
Academic Training							
BA/BSc				13		13	
AQ Courses				8		7	
Principal Course				1		1	
MA/Med				1		1	
Conferences				11		9	
Summer Institutes				6		6	

Table 2: Internal Reliability of Student Variables (N=290)

	Number of Items	Range	Cronbach's Alpha
Self Efficacy	6	1-10	.85
Self-Evaluation	6	1-10	.84
Self-Evaluation Attitudes	10	1-5	.75
Goal Orientations			
mastery	9	1-5	.84
ego	3	1-5	.62
affiliative	3	1-5	.54

Table 3: Unadjusted Means and Standard Deviations of Student Variables, by Experimental Condition

	Treatment (N=148)		Control (N=148)	
	M	SD	M	SD
Self-Efficacy	7.40	1.58	7.28	1.88
Self-Evaluation	7.34	1.69	7.03	1.66
Self-Evaluation Attitudes	3.85	.63	3.97	.62
Goal Orientations				
mastery	3.77	.75	3.88	.76
ego	2.92	.96	2.98	.95
affiliative	3.21	1.04	3.17	1.16
Achievement				
pre	3.62	1.37	2.90	1.22
post	4.17	1.23	3.23	1.31
	Percent		Percent	
Accurate				
pre	35		26	
post	64		40	
Gender				
male	54		51	
female	46		49	
Age				
9 or under	23		23	
10	37		40	
11	36		29	
12 & over	5		8	

Table 4: Posttest Accuracy of Underestimating, Accurate, and Overestimating Students, by Experimental Condition

Pretest Accuracy	% Accurate on Posttest
Underestimate (N=5)	
treatment (N=4)	50
control (N=1)	0
Accurate (N=90)	
treatment (N=53)	72
control (N=37)	54
Overestimate (N=202)	
treatment (N=91)	30
control (N=111)	18

Table 5: Effect of Self-Evaluation on Student Achievement: Results of Analysis of Covariance

Source of Variation	SS	DF	MS	F	p	Partial Eta Sqd
Within + residual	3.35	291	.01			
Pretest	1.30	1	1.30	112.98	<.001	.280
Self-Evaluation Accuracy	.01	1	.01	1.30	.255	.004
Treatment	.08	1	.08	7.22	.008	.024
Pretest x Treatment	.11	1	.11	9.50	.002	.032
Self-Evaluation x Treatment	.00	1	.00	.03	.859	.000
Model	2.50	5	.50	43.44	<.001	
Total	5.86	296	.02			
R-squared .427						

Table 6: Adjusted Posttest Achievement Means and Standard Deviations of Low and High Achieving Groups, by Experimental Conditions

	Low Pretest Achievement		High Pretest Achievement	
	Treatment (N=73)	Control (N=109)	Treatment (N=75)	Control (N=40)
Pretest Mean	.53	.51	.76	.73
Pretest SD	.12	.12	.01	.01
Posttest Mean	.65	.56	.75	.71
Posttest SD	.11	.14	.01	.11

Appendix: Assessment Scales for Junior Division Narrative

Level 6

The work demonstrates a confident command and integration of all the elements of writing. The content is often strikingly creative and imaginative (e.g., evidence of risk taking).

Possible Characteristics

- The controlling idea and its development are insightful and original, and consistent with the narrative form.
- The organization is subtle; the control is secure; the style reinforces the purpose; events are well sequenced with supporting detail.
- The voice is confident; there is a sense of engagement with the topic and an effective relationship with the audience; voice is appropriate to the narrative words.
- The control of written conventions of language is skilful; rare errors in spelling and minor errors in grammar and punctuation may exist but do not affect the overall impact; they may be the result of the difficulty of the writing task and/or risks taken by the student.

Level 5+

Possible Characteristics

- Writer is in command of elements of narrative but not completely.
- Controlling idea original and creative but not striking.
- Development of ideas demonstrate originality.
- Voice is clear and effective.
- Strong sense of reader.
- Effective level of word choice.
- Very few errors in conventions; spelling errors a result of use of difficult word choice.

Level 5

The work shows an effective control and integration of all the elements of writing. The content is thoughtful and thorough.

Possible Characteristics

- The controlling idea and its development are thoughtful and thorough, and consistent with the narrative form.
- The organization is effective; the style is appropriate to the purpose and the narrative form.
- The voice is clear; there is a strong sense of audience.
- The control of the written conventions is sound; any errors in spelling, grammar, and punctuation do not detract from the overall impact.

Adapted from Ontario Ministry of Education and Training Standards

Level 4+

Possible Characteristics

- Good control of the elements of narrative writing.
- Some evidence of originality.
- Consistent narrative voice; good awareness of audience.
- Organization is clearly evident.
- Good evidence of style.
- Conventions in good control; there may be some errors but they do not detract from meaning.

Level 4

The work shows control of the elements of writing. It is generally integrated. The content is clear and complete.

Possible Characteristics

- The controlling idea and its development are clear but may be conventional or derivative (e.g., a summary of events).
- Organization is capable; there is a clear attempt to connect style and purpose with narrative form.
- The voice is apparent but may fluctuate; there is an awareness of audience.

The control of the written conventions is capable; infrequent errors may detract from the overall impact of the work but do not affect the meaning.

Level 3+

Possible Characteristics

- Writer makes an obvious effort to involve the reader.
- Elements of narrative writing are under control but any lack of control can effect meaning.
- Integration of elements, development of story is almost complete.
- Organization is apparent but still not really clear; paragraphing is used.
- Evidence of a narrative voice, but not consistent.
- Control of conventions is capable but still can impact on meaning.

Level 3

The work shows control of most of the elements of narrative writing. Some integration is apparent. The content may be simple or unoriginal.

Possible characteristics

- The controlling idea and its development are apparent and show some balance or consistency; ideas convey surface meaning.
- Organization is apparent; there is some attempt to connect style and purpose.
- There is a sense of voice with some control; there is an occasional awareness of audience.
- Control of the written conventions of language is evident; errors occasionally detract from the impact and the meaning.

Level 2+

Possible Characteristics

- A firm grasp of the basic elements (conventions, sentence structure--not necessarily paragraphs).
- Limited sophistication/maturity of ideas.
- Controlling idea is apparent but uneven.
- Some organization is apparent, but little or no attempt to connect style to purpose/theme.
- Narrative voice emerging but distinction between writer's personal voice and the narrative voice not clear (writer-oriented text vs. reader-oriented text).
- Conventions distract but understanding of ideas is possible.

Level 2

The work shows grasp of some of the basic elements of narrative writing; the writing conveys simple ideas.

Possible Characteristics

- The controlling idea and its development are limited but discernible; ideas are superficial.
- Organization is attempted; style is simple and unconnected to the purpose.
- Voice may be often limited to a personal, vernacular register; awareness of audience is limited or absent.
- Grasp of the written conventions of language is tentative; errors are distracting and often interfere with the reader's understanding of the ideas.

Level 1

The work shows a minimal grasp of some of the basic elements of writing. The content conveys unconnected or fragmented ideas.

Possible Characteristics

- The writing expresses some unconnected ideas, but no discernible controlling idea.
- Organization is not discernible.
- Voice is limited to personal, vernacular register; awareness of audience is absent.
- Grasp of the written conventions of language is minimal; errors impede expression and comprehension.

Additional Scoring Notes

If the work was less than half a page, the passage was scored no higher than 2.

BEST COPY AVAILABLE